MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190498

Extending Achilles Heel Data Quality Tool with New Rules Informed by Multi-Site Data Quality Comparison

Vojtech Huser[a], Xiaochun Li[b], Zuoyi Zhang[b], Sungjae Jung[c], Rae Woong Park[c], Juan Banda[d], Hanieh Razzaghi[e], Ajit Londhe[f], Karthik Natarajan[g]

[a] National Library of Medicine, MD, USA, [b] Indiana University, IN, USA, [c] Ajou University School of Medicine, Suwon, South Korea, [d] Georgia State University, GA, USA, [e] Children's Hospital of Philadelphia, PA, USA, [f] Janssen Research & Development, NJ, USA, [g] Columbia University, NY, USA

Abstract

Large healthcare datasets of Electronic Health Record data became indispensable in clinical research. Data quality in such datasets recently became a focus of many distributed research networks. Despite the fact that data quality is specific to a given research question, many existing data quality platform prove that general data quality assessment on dataset level (given a spectrum of research questions) is possible and highly requested by researchers. We present comparison of 12 datasets and extension of Achilles Heel data quality software tool with new rules and data characterization measures.

Keywords: *data quality, observational study*

Introduction

Data quality is an important pre-requisite for research on Electronic Health Record (EHR) data. In recent years, several efforts and tools emerged that perform data quality assessment (DQA).[1] Another important trend is that research is increasingly conducted using distributed research networks. Such networks often provide tools to their data partners that lower the barrier to join or participate within the network and help with data preparation or analysis execution.

The Achilles tool from the Observational Health Data Sciences and Informatics Consortium (OHDSI) is one such tool that performs data characterization and includes an Achilles Heel part that contains rules for checking data quality (DQ). The Achilles tool has been first deployed in October 2014 (version 1.0) with several updates (versions 1.1 to 1.6) during a period from 2014 to 2018. Since 2016, the web-based user interface part of Achilles was incorporated into the OHDSI Atlas tool, which is a new interface that integrates into one interface several previously developed OHDSI tools.

In developing the Achilles tool, the OHDSI consortium actively encourages researchers to submit requests for new data quality checks or insightful data visualizations that would extend the tool's utility. The Achilles' software repository receives numerous inputs (in a form of Github issues) that identify such new DQ measures or checks. In addition to this ongoing feedback, European EMIF research network conducted a formal survey of the tool that indicated the need for new features.

This study describes a set of extensions of the Achilles tool based on a comparison of data quality indicators of several healthcare datasets.

Methods

The study had two goals. The first goal was to compare data quality characteristics across datasets. Informed by this comparison, the second goal was to extend Achilles with new features and new data quality rules that would improve the assessment of data quality generated by the tool. This study includes a larger set of exported dataset metadata compared with a previous study done by our team, that only focused on Achilles Heel output messages.

The Achilles tool currently generates over 170 measures. However, many healthcare dataset administrators are not permitted to share such comprehensive set of dataset indicators. To be able to conduct our comparison, we designed a smaller set of measures generated by Achilles pre-computations that includes only measures that were deemed acceptable by the dataset administrators.

To maintain a data aggregation privacy-preserving principle for our comparison, our study used a small-cell count threshold of 11+ patients per aggregated count. Achilles tool allows suppressing pre-computations that result in small counts of patients (or small counts of providers, or healthcare events). This filtering is done either when Achilles pre-computations are executed, but if it was not done during the Achilles pre-computation phase, our methodology enforced it again during when site data extract generation. The R package for our study (called DataQuality) is open source and available on the Github platform at https://github.com/OHDSI/StudyProtocolSandbox/tree/master/DataQuality.Actual input data for the study consisted of the following: (1) subset of Achilles analyses converted to ratios (for example, ratio of persons with at least one visit by visit type); (2) Achilles derived measures (for example, percentage of unmapped source data concepts by domain) and (3) an approximate size of the dataset (for example, <10k, 10-99k,100k-1M, 1-5M,5-9M and >10M; exact size of populations is masked into a dataset size category). Sample input data (for a synthetic SynPuf OMOP dataset) is available at Github.

Each dataset was assigned a meaningless identifier to facilitate the comparison. The purpose for this dataset masking is the fact that data quality comparisons can lead to withdrawal of a data partner from a research consortium (or an analysis project) if a particular partner's dataset is identified as having low quality data. Masking was done to avoid this outcome and to focus on advancing the methodologies for DQA. For the same reason, neither a list of individual datasets is provided. We plan to

destroy individual site aggregated data used as input at 6 months after the publication of the study results. To protect the sites, only masked and isolated combined comparisons are reported in this article. Non-aggregated, single dataset DQ data are never posted publically.

Determination of goodness of fit or "data fitness" is highly dependent on the research question being asked. This phenomenon was described earlier and is sometimes referred to a task-dependence nature of DQA.[2] A dataset that only contains inpatient events and data may not be appropriate for general research questions (e.g., descriptive study of a course of a disease); however, it may be sufficient for a subset of research questions (e.g., inpatient-only research questions).

One can conclude that without knowing the specific research question context, any data quality assessment is impossible to pre-empt. This requirement for specifying research question context up-front makes development of general DQA tools almost impossible. However, existing DQA tools and efforts indicate that some general DQA rules indeed exist.

To partially overcome this problem ("data fitness for what?"), we assumed that the dataset being assessed represents lifetime record of general population and the tool should perform DQA for a wide range of possible research questions ("*general* data fitness for a wide *range of research questions*"). Once a general DQA analysis is done, a researcher with a specific research analysis can simply ignore DQA messages that do not apply to his/her context. (e.g., ignore messages about lack of eye doctor's visit and eye care data if data about vision care are not essential for his/her research question).

Results

A total of 12 datasets were compared in the study; however due to use of prior Achilles versions by some sites, comparison of some newly implemented measures are made on data from datasets that implemented at least Achilles version 1.4 at the time of our study data extraction.

Version 1.6 of Achilles contains a total of 44 data quality rules (also called data quality checks). A total of 12 rules are *model conformance rules* that check adherence to the CDM specification. For example, a model conformance rule may require that provider specialty column contains only concepts that are indeed specialties. The remaining 22 rules are *data quality rules* that check for data completeness, data plausibility or other data quality problems. Such rules can be considered model-independent and should be portable to other data models, such as Sentinel model or PCORNet. The pooled dataset of all Achilles Heel messages from all datasets consisted of 546 messages. Median number of Heel messages for a single dataset was 51 with a median of 25 for errors, 22 for warnings and 4 for notification. Poster will show evaluation of severity of each rule violation by computing median record counts for each rule. The second goal of our study was to add new functionality (either new DQ rules or new DQ measures to Achilles) based on availability of data about multiple CDM datasets. The results are divided into multiple sections according to the data domain of the new rule and will be included in the poster.

(1) **Empirical rules:** Comparing selected dataset parameters and computing 90^{th} or 10^{th} percentile and using them as benchmark thresholds.

(2) **Data density rules:** We considered data density at three levels (*concepts per person* as a number of distinct measurements per person (e.g., count of 2 measurements per person, such as cholesterol and hematocrit). This comparison aims at "data breadth"; *records per person* as total number of all measurement records per person (e.g., count of 8 tests, such as 3 LDL cholesterol and 5 hematocrit measurements). This comparison aims at "data depth"; records per visit as a data density measure on a visit level. Because visits with no measurements occur, the per visit ratio measure can be below 1. However, for a ratio looking at clinical notes (if in scope for the dataset), it may be reasonable to expect at least one note per visit.

(3) **Minimum-data patients:** For many research question, at least one data point in a given clinical data domain (such as medications) is required for any meaningful analysis. For example, for analyzing event prevalence, using a proper denominator and determining the size of the relevant population can significantly affect the reported measure. We determined empiric thresholds for existing Achilles DQA measures that count number of patients with at least one event in a clinical data domain. (e.g., patients with at least one visit, patients with at least 1 diagnosis and 1 medication.

(4) **Unmapped data:** OMOP CDM allows storage of data that is not fully semantically mapped to standard concepts (for example, drug exposure data may include data rows that have a value of 0 ('No matching concept') in drug_concept_id while the yet-to-be-mapped local code is stored in drug_source_value). We introduced measures computing unmapped data and threshold rules for several domains, such as Conditions, Procedures or Drug Exposure.

Discussion and Conclusion

Our current method for picking an empiric threshold is using a fixed threshold (e.g., 10^{th} percentile). Future methodology revision may alter this approach for each considered DQ measure. Another limitation is our primary focus on OMOP CDM sites. Our extension to Achilles rule knowledge base, however, point to what data measures are required by each rule and whether a rule is terminology dependent. We compared data quality indicators across several datasets. We arrived at empirical values that could be used as thresholds for several DQA measures. The study resulted in several new data quality checks being added to Achilles.

We would like to thank Ritu Khare, Taha Abdul, Chris Knoll and Martijn Schuemie. VH work was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC).

References

[1] H. Estiri and K. Stephens, DQe-v: A Database-Agnostic Framework for Exploring Variability in Electronic Health Record Data Across Time and Site Location. , eGEMs (Generating Evidence & Methods to improve patient outcomes) 1 (2017).

[2] N.G. Weiskopf and C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J Am Med Inform Assoc 20 (2013), 144-151.

Address for correspondence

vojtech.huser@nih.gov